J Grid Computing
https://doi.org/10.1007/s10723-018-9469-8

Special Issue on Big Data and Digital Transformation

José Luís Pereira · Orlando Belo · Pascal Ravesteijn

Received: 9 October 2018 / Accepted: 9 October 2018

The world we live in is more and more of a digital nature. Nowadays, more than ever, we are witnessing a fundamental trend - technological innovations are continuously appearing and rapidly made available to everyone. These have a huge impact on people, who, for the most, easily adhere to and use them, not only in their personal lives, but also at work. At the enterprise level, easy access to global markets and new ways of working and conducting business becomes possible due to technological innovations, leading to innovative business models. Similarly, governments, municipalities and public organizations are developing new ways to reach the citizens and interact with them, in manners that were impossible to foresee a few years ago. All these changes, fueled by the continuous stream of technological innovations, are widely known as *Digital Transformation*.

Digital Transformation embraces all society being often the cause of disruptive changes. Indeed, from smart cities and urban computing to virtual enterprises and the cloud; from social media and virtual and augmented reality to the Internet of Things (IoT), just to name a few modern technological buzzwords, our lives are progressively being more and more digitalized. One of the major consequences of this unstoppable movement to a more digital world is the creation of increasingly larger volumes of data, which need to be properly collected and managed.

The huge quantities of data that are being produced (*Volume*), the increasingly higher paces at which these data are generated (*Velocity*) and the multitude of sources from which data are collected (*Variety*) require different and innovative data management technologies, in order to cope with these new challenges. Furthermore, in the era of Digital Transformation, the flexibility and distributed nature of the data processing solutions are the norm. These are some of the main reasons behind the so-called *Big Data* momentum.

Facing the new challenges in terms of data generation, brought by the Digital Transformation movement, Big Data answered by offering new data management technologies, capable of handling the higher Volume, Velocity and Variety characteristics of the new data. Therefore, Digital Transformation and Big Data solutions are deeply related - one cannot happen without the other.

For this special issue on Big Data and Digital Transformation, we have collected five articles that constitute relevant developments and advances in the area. These articles were selected and improved

J. L. Pereira (✉)
University of Minho, Guimarães, Portugal
e-mail: jlmp@dsi.uminho.pt

O. Belo
University of Minho, Braga, Portugal
e-mail: obelo@di.uminho.pt

P. Ravesteijn
HU University of Applied Sciences Utrecht, Utrecht, The Netherlands
e-mail: pascal.ravesteijn@hu.nl

Published online: 24 October 2018

during a two-round review process, in accordance with the standard practices of the Journal of Grid Computing. The resulting five contributions cover some key aspects and developments regarding Big Data technologies in the realm of Digital Transformation.

The first article, "A Task-Based Greedy Scheduling Algorithm for Minimizing Energy of MapReduce Jobs", by Mostafa Yousefi and Maziar Goudarzi, addresses a very relevant issue, which is of utmost importance in the Big Data era: the reduction of energy consumption during the parallel processing of very large data jobs. Taking into consideration that the need to process even larger data jobs will increase in the future, the importance of reducing the energy-related operational costs is obvious because of the growing environmental impact (carbon footprint). In this article, authors present TGSAVE – a greedy scheduling algorithm that minimizes total energy consumption of MapReduce jobs for Big Data applications in heterogeneous environments. According to their experiments with several benchmarks, TGSAVE stands as a substantial improvement to an already existing energy aware scheduler for MapReduce jobs – EMRSA. Besides minimizing energy consumption, TGSAVE assures the satisfaction of Service Level Agreements (SLAs) without significant performance losses.

As Preeti Gupta, Rajni Jindal and Arun Sharma state in the second article "Community Trolling: An Active Learning Approach For Topic Based Community Detection In Big Data", in the age of Big Data manual inspection of huge datasets is simply not possible, therefore semi-automatic data filtering approaches and data classification techniques are required. In this article, using as a motivating example the huge amounts of data produced by people during their interactions in social networks, the authors demonstrate the viability of active learning to develop an approach for topic based community detection in big datasets. They termed the active learning phase as community trolling, acting as a precursor to community detection, which plays an important role in social network analysis. Although social media data might contain many noisy data, community trolling selectively samples the data relevant to the current context using active learning. In this article, the effectiveness of the proposed approach was demonstrated by using it on a real world Tumblr social network dataset.

The goal of the third article "Incentive Mechanisms for Resource Scaling-out Game of Stream Big Data Analytics", authored by Xiaoyuan Fu, Jingyu Wang, Qi Qi, Jianxin Liao and Tonghong Li, is to present a solution to the problem of resource scaling-out in the context of stream-processing Big Data analytics. Stream-processing has emerged as a way to tackle the data that enters the system continuously and, particularly in the case of Big Data analytics, there is always the need to scale-out resources when input data increases suddenly. Typically, the way to scale-out the available resources for a job is to increase the parallelism degree of the platform. In this article, authors have used the Apache Storm stream-processing platform to develop a new approach in which a game model and incentive mechanisms are used to tackle the resources scaling-out for steep-increasing input data of real-time applications. They claim their work can be applied to stream-processing platforms in general, to make decisions for improving the platforms parallelism.

The fourth article, "Big Data-oriented PaaS architecture with disk-as-a-resource capability and container-based virtualization", authored by Jonatan Enes, Javier López Cacheiro, Roberto Rey Expósito and Juan Touriño deals with the lack of proper support that Big Data applications get from the most commonly used cloud providers. According to authors, in order to efficiently execute Big Data applications, it is vital that cloud providers change the way hardware infrastructure resources are managed, paying particular attention to storage resources. The main problem is the use of shared disks in local computing nodes, which severely reduces the performance of Big Data applications. In this article, a new Platform-as-a-Service (PaaS) architecture is proposed, specifically oriented to support the requirements of Big Data applications, which allows for scheduling disks as resources, providing a disk-as-a-resource capability. Differently from standard cloud and virtualization services (Azure, AWS, OpenStack, etc.), which assume sharing and virtualization of storage, in the proposed PaaS architecture the local disks at each computing node can be exposed as resources. The performance of the proposed PaaS architecture has been compared with the common OpenStack framework running representative Big Data workloads. The results shown significant performance improvements.

In the fifth article, "A Dynamic Spark-Based Classification Framework for Imbalanced Big Data", authors Nahla B. Abdel-Hamid, Sally ElGhamrawy, Ali El Desouky and Hesham Arafatt make a relevant contribution to the problem of classification of imbalanced Big Data. Organizations have concluded that their large volumes of data are only as valuable as the insights that might be extracted from them. They need appropriate tools, such as data classification techniques, to extract the hidden value embedded in their data. Recently, classification of imbalanced data has become a relevant research topic as many Big Data applications may suffer from the imbalance dataset problem. In this article, using the Apache Spark (a widely known platform for carrying out distributed operations on Big Data), a Spark-Based Mining Framework (SBMF) is proposed to classify imbalanced Big Data. The SBMF was tested and its performance assessed using datasets with different imbalanced ratios and distinct sizes, from moderate size imbalanced datasets, to large imbalanced datasets, to reflect Big Data scenarios. The results obtained, comparatively to other recent proposals, were very encouraging.

After a thorough and rigorous review process, we believe that these five articles constitute an interesting set of contributions addressing the theme of Big Data and offering solutions in order to enable Digital Transformation. We would like to thank the many authors who have submitted their work to this special issue, for their valuable insights and contributions. A special thanks also to the reviewers who, with their commitment and expertise, have helped us to select the best articles and to improve their content. We hope that readers find this special issue an interesting and valuable source of information to their own work. Finally, we would like to express our gratitude to the Editor-in-Chief of the Journal of Grid Computing, Prof. Peter Kacsuk, for giving us the opportunity to organize this special issue and for his continuous support during the all process.